DOCUMENT RESUME

ED 269 420 TM 860 231

AUTHOR Zwick, Rebecca

TITLE Assessment of the Dimensionality of NAEP Year 15 Reading Data.

INSTITUTION Educational Testing Service, Princeton, N.J.
SPONS AGENCY National Inst. of Education (ED), Washington, DC.
REPORT NO ETS-RR-86-4
PUB DATE Jan 86
GRANT NIE-G-83-0011
NOTE 55p.
PUB TYPE Reports - Research/Technical (143)

EDRS PRICE MF01/PC03 Plus Postage.
DESCRIPTORS *Correlation; Educational Assessment; Elementary Secondary Education; *Factor Analysis; Grade 4; Grade 8; Grade 11; Item Analysis; *Latent Trait Theory; Mathematical Models; Reading Achievement; *Reading Tests; *Scaling; Statistical Studies
IDENTIFIERS Guttman Scales; *National Assessment of Educational Progress; Phi Coefficient; Tetrachoric Correlation; *Unidimensionality (Tests)

ABSTRACT

Reading test data from the National Assessment of Educational Progress (NAEP) were scaled using the unidimensional item response theory model. Data were collected for students aged 9, 13, and 17. To determine whether the responses to the reading items were consistent with unidimensionality, four different methods were applied: (1) principal component analysis of phi and tetrachoric correlation matrices; (2) principal component analysis of the image correlation matrix, a method based on the work of Guttman; (3) R. D. Bock's full-information factor analysis; and (4) P. R. Rosenbaum's test of unidimensionality, monotonicity, and conditional independence. Balanced incomplete block (BIB) spiralling was used with this year's NAEP to assign test items to booklets. This permitted the estimation of inter-item correlations, but resulted in an unusual pattern of missing data. Results from the analyses conducted for each age group were different from the analysis of the 25 items administered in all three samples. It was concluded that it was not unreasonable to regard the reading items as measures of a single dimension of reading proficiency. (Author/GDC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

RR-86-4

ED269420

RESEARCH REPORT

ASSESSMENT OF THE DIMENSIONALITY OF NAEP YEAR 15 READING DATA

Rebecca Zwick

U.S. DEPARTMENT OF EDUCATION
NATIONAL INSTITUTE OF EDUCATION
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

☑ The document has been reproduced as received from the person or organization originating it.

☐ Minor changes have been made to improve reproduction quality.

• Points of view or opinions stated in this document do not necessarily represent official NIE position or policy.

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY

H. Weidenmiller

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Educational Testing Service
Princeton, New Jersey
January 1986

TM 860 231

Assessment of the Dimensionality of NAEP Year 15 Reading Data

Rebecca Zwick

Psychometrics Research Group

Educational Testing Service

January 15, 1985

Acknowledgments

The author is grateful to the many colleagues who provided consultation on dimensionality assessment methods and offered comments on earlier drafts of this report. They include Al Beaton, Darrell Bock, Neil Dorans, Paul Holland, Bob Mislevy, Paul Rosenbaum, and Ledyard Tucker. Thanks also to Dick Harrison, Bruce Kaplan, and Dorothy Thayer, who programmed the procedures described in this report, Natalie Roca, who conducted analyses and reviewed literature, and Marilyn Brennan, who prepared the manuscript. The work upon which this publication is based was performed pursuant to Grant No. NIE-G-83-0011 of the National Institute of Education. It does not, however, necessarily reflect the views of that agency.

Abstract

The reading data from the 1983-1984 NAEP survey were scaled using a unidimensional item response theory model. To determine whether the responses to the reading items were consistent with unidimensionality, four methods were applied: principal component analysis of phi and tetrachoric correlation matrices; principal component analysis of the image correlation matrix, a method based on the work of Guttman (1953); Bock's full-information factor analysis (Bock, Gibbons, and Muraki, 1985); and Rosenbaum's (1984a) test of unidimensionality, monotonicity, and conditional independence. Results indicated that it was not unreasonable to regard the reading items as measures of a single dimension.

Table of Contents

List of Tables iii

1. The National Assessment of Educational Progress. 1

1.1 The unidimensionality assumption in item response theory . .1

1.2 Robustness of IRT estimation procedures 2

2. Methods of dimensionality assessment for dichotomous data. . . . 3

3. Methods used to assess dimensionality of NAEP reading data . . . 9

3.1 Properties of NAEP data 10

3.1.1 Items included in dimensionality analyses. 10

3.1.2 Missing data pattern 13

3.2 Principal components analyses of inter-item correlation matrices. 15

3.2.1 Application of guessing corrections to tetrachoric correlations. 16

3.3 Principal components analysis of the image correlation matrix 20

3.4 Bock's full information factor analysis 28

3.5 Rosenbaum's test of unidimensionality, monotonicity, and conditional independence. 33

3.5.1 Across-grage analyses. 37

4. Conclusions. 39

References 41

Appendix: A procedure for obtaining a Gramian matrix that approximates a BIB correlation matrix for NAEP items. 47

List of Tables

1. Number of Items and Students Available for Dimensionality Analyses 11

2. Eigenvalues and Descrptive Statistics for Phi Matrices. .17

3. Eigenvalues and Descriptive Statistics for Tetrachoric Matrices 18

4. Eigenvalues of the Image Correlation Matrix 24

5. First Five Eigenvalues of Correlation and Image Correlation Matrices for Simulation Data (30 items with NAEP item parameters). 27

6. Results of Rosenbaum Analyses 36

1. The National Assessment of Educational Progress

The National Assessment of Educational Progress (NAEP) is a congressionally mandated survey of the educational achievement of American students that has been conducted since 1969. Educational Testing Service assumed responsibility for it in 1983. During the 1983-1984 academic year (Year 15 of NAEP), ETS collected data on three so-called grages: 9/IV, 13/VIII, and 17/IX. In NAEP parlance, a grage is the union of an age, denoted by an Arabic numeral, and a grade, denoted by a Roman numeral.

The subject areas assessed during Year 15 were reading and writing. Only the reading items are discussed in the present report.

1.1 The unidimensionality assumption in item response theory

In order to determine whether it was reasonable to regard the reading items administered in the Year 15 NAEP data collection as measures of a single construct, a series of analyses of the dimensionality of the reading data was performed. Dimensionality analyses were conducted both within and across the three grages, 9/IV, 13/VIII, and 17/IX. It was important to investigate the dimensionality issue because the validity of the item response theory (IRT) model used to estimate reading proficiency in the 1983-1984 NAEP survey rests on the assumption of unidimensionality. It should be noted, however, that regardless of whether an IRT model is used, it is ordinarily assumed that items on an achievement test can be treated as measures of a single dimension, in this case, reading proficiency. Scoring a test by simply summing the item scores involves an implicit assumption of unidimensionality; IRT scaling formalizes this assumption.

The reading data were analyzed using the three-parameter logistic model (Birnbaum, 1968; Lord, 1980) in which P_{ij}, the probability that subject i gets item j correct can be expressed as follows:

$$P_{ij} \equiv P(x_{ij} = 1|\theta) = c_j + \frac{1 - c_j}{1 + e^{-1.7a_j(\theta_i - b_j)}} \qquad [1]$$

where θ_i is the proficiency parameter for person i, a_j is the item discrimination parameter, b_j is the item difficulty, and c_j can be interpreted as the probability that a person with very low ability gets item j correct. (Model parameters were estimated using BILOG [Mislevy and Bock, 1982]; details are provided in a separate report on scaling.) In applying a model of this kind, it is assumed that the only examinee characteristic that affects item response is a single latent variable, θ.

1.2 Robustness of IRT Estimation Procedures

In practice, the assumption of unidimensionality, required for the application of conventional IRT models, will always be violated to some degree. In order to make a more objective determination as to what constitutes an important departure from unidimensionality, we need to know more about the robustness of the IRT estimation procedures to violations of the unidimensionality assumption. Unfortunately, little work has been done in this area. Reckase (1979) and Drasgow and Parsons (1983) investigated the results of estimating the three-parameter logistic model, using LOGIST (M. S. Wingersky, 1983), under violations of the unidimensionality assumption. (The one-parameter and two-parameter logistic models were also examined by Reckase,

1979, and Drasgow and Parsons, 1983, respectively.) Reckase's study was based on five actual data sets and five data sets constructed to have specific factor structures. He concluded that LOGIST estimates "the first principal component when it is large relative to other factors .... good ability estimates can be obtained ... even when the first factor accounts for less than 10 percent of the test variance, although item calibration results will be unstable. For acceptable calibration, the first factor should account for at least 20 percent of the test variance" (p. 228). Drasgow and Parsons (1983) made use of a hierarchical model with a general latent trait as well as five group factors to simulate various kinds of latent structures. One of their conclusions was that, in the simulated data designed to resemble "moderately heterogeneous achievement tests and attitude assessment instruments" (p. 193), LOGIST still recovered the latent trait and provided acceptable estimates of the item parameters (p. 198). There is no reason to believe that the effects of multidimensionality on BILOG (Mislevy and Bock, 1982), which was used to scale the NAEP data, would differ from the results obtained with LOGIST (Mislevy, personal communication, October, 1985). These findings suggest that IRT scaling procedures can produce satisfactory results under moderate departures from unidimensionality.

2. Methods of dimensionality assessment for dichotomous data

The traditional psychometric approach to the assessment of dimensionality is through factor-analytic methods. Factor analysis often produces satisfactory results when each of the variables is the score on a multi-item test. When each of the measures is the response to a dichotomously scored item, however, it is now well known that linear factor analysis of Pearson (phi)

correlations does not, in general, yield a correct representation of the dimensionality of the item pool (see, e.g., Carroll, 1945, 1983; Hulin, Drasgow, and Parsons, 1983; McDonald and Ahlawat, 1974, Mislevy, in press). The fundamental problem is that in computing phi correlations, item responses are treated as true dichotomies. In applying a linear factor analysis model, we are hypothesizing that dichotomous variables are linear combinations of continuous latent variables with infinite range, a mathematical impossibility. In fact, the regression of a dichotomous item on a continuous latent variable must be nonlinear. The best linear approximation to the nonlinear regression will depend on the region in which the data are most dense (Mislevy, in press); that is, it will be related to the item mean, or difficulty (as defined in classical test theory). From this perspective, it is not surprising that linear factor analysis of dichotomous items often produces a second factor, typically called a difficulty factor, that is related to item difficulty, but appears to be unrelated to any substantive property of the items. There can, in fact, be more than one such spurious factor (as is the case for items that form a perfect Guttman scale), but ordinarily, only one is substantial in size.

A related problem with the phi coefficient, which can be regarded as another manifestation of the departure from the assumptions of classical factor analysis, is that its magnitude is determined in part by the relative values of the means of the two variables, which in this case are the item difficulties. Regardless of the underlying relationship between the items, the phi coefficient can reach unity only if the two items have identical proportions correct.

As an alternative to phi coefficients, tetrachoric correlations between items can be obtained. In computing tetrachorics, it is assumed that the item responses are functions of underlying continuous variables that have a bivariate normal distribution. The model dictates that, for each item, individuals who have values greater than a certain threshold on the underlying response variable get that item correct; individuals with values lower than the threshold get it wrong. Using the bivariate normality asssumption, the correlation between the unobserved continuous variables can be inferred from the 2 x 2 table of item responses. Of course, tetrachoric correlations do not provide a valid measure of association if bivariate normality does not hold. Furthermore, the occurrence of guessing violates the above model, which postulates that the probability that an individual gets an item right is a function only of his value on the underlying response variable. When guessing does occur, factor analysis of tetrachorics can produce spurious factors (see Carroll, 1945, 1983; Hulin, Drasgow, and Parsons, 1983). Adjustments for guessing are theoretically possible, but often lead to unacceptable results in practice. (Attempts to adjust for the effects of guessing in the NAEP analyses are discussed in section 3.2.1.) Additional problems are inaccuracies in the computation of tetrachorics as they approach +1 or -1, the large standard errors of the coefficients, and the occurrence of non-Gramian matrices of sample tetrachorics, even when data are complete. (In the case of the NAEP analyses, in which a large proportion of data are missing by design, the negative eigenvalues tend to comprise a large proportion of the trace of the tetrachoric matrix; see section 3.1.2 and Table 3.)

It is clear that conventional factor analysis of phi and tetrachoric correlations is not a satisfactory means of investigating dimensionality. Unfortunately, no uniformly accepted statistical procedures for dimensionality assessment exist for the case of dichotomous variables. As a result, a vast literature on the subject has developed, particularly during the last ten years, as the use of IRT models has increased. Some methods which have gained attention recently are briefly described here; more detailed reviews of dimensionality assessment are given by Hattie (1984, 1985), Hulin, Drasgow, and Parsons (1983, Chapter 8), and Mislevy (in press).

Factor-analytic methods that have been proposed to overcome the problems described above include factor analysis of item parcels, nonlinear factor analysis, the generalized least squares methods developed by Christofferson (1975) and Muthén (1978) and the full-information maximum likelihood method of Bock (Bock, Gibbons, and Muraki, 1985).

Factor analysis of item parcels is achieved by grouping items into meaningful subtests (the so-called parcels) and then applying conventional factor-analytic methods to the parcel scores. This method was applied by Cook and Eignor (1984) to a portion of the NAEP data collected in 1979-1980 and by Cook, Eignor, Dorans, and Petersen (1985) to SAT data. One practical problem with this approach is that it may be difficult to classify certain items a priori. Furthermore, if the item parcels differ in average difficulty, the obtained factor structure may be influenced to an undesirable degree by item difficulty, as in the dichotomous case (Kingston and Dorans, 1982). A more fundamental drawback is that this approach does not assess directly the properties of individual items. Because item scores do not enter the

analysis, it is possible for items that measure a property other than the one of interest to go undetected. Finally, the application of this approach to the complete NAEP data set is virtually ruled out because examinees do not all receive the same items (see section 3.1). (The Cook and Eignor [1984] analysis was based on a subset of examinees who had been administered the same items.)

In a series of publications, McDonald presented a theory of nonlinear factor analysis (e.g., McDonald, 1967, 1983). In McDonald's model, $P(x_{ij} = 1 \mid \underline{\theta})$, the conditional probability that an examinee answers an item correctly, given his observed vector of latent traits, $\underline{\theta}$, is expressed as a nonlinear function of the latent traits. For example, in one version of the model, $P(x_{ij} = 1 \mid \underline{\theta})$ is expressed as a weighted sum of polynomial functions of the latent traits. Simulation studies of the effectiveness of nonlinear factor analysis as a method of dimensionality assessment have led to inconsistent findings. Hambleton and Rovinelli (in press) found that a one-factor polynomial model with linear and quadratic terms provided a good fit to a simulated unidimensional data set, unlike a one-factor linear model. Furthermore, a two-factor polynomial model provided a good fit to two-dimensional simulated data. Based on this and other findings, Hambleton and Rovinelli concluded that nonlinear factor analysis is one of the most promising methods for assessing the dimensionality of dichotomous data. On the other hand, Hattie (1984) concluded that the sum of absolute residual covariances from nonlinear factor analysis was not an effective index of dimensionality because results from the unidimensional and multidimensional data sets were not sufficiently distinct.

Christofferson (1975) developed a factor-analytic method for dichotomous data that involves expressing the expected proportion correct for each item and for the joint proportions correct for each pair of items as a function of item thresholds (see above and section 3.4, below) and factor loadings. The weighted distance between the observed and modeled values of these proportions is then minimized using generalized least squares (GLS) methods. Christofferson's solution makes use of the information contained in the three-and four-way margins of the n-way contingency table of item responses (see Christofferson, 1975, Appendix 2; Mislevy, in press), unlike conventional factor analysis of phi or tetrachoric correlations, which makes use of only the one-and two-way marginals. Solving for estimates of the thresholds and loadings requires numerical integration and is therefore computationally burdensome. Muthén (1978) developed an alternative GLS method that reduces the computational requirements to some degree. However, application of both Christofferson's and Muthén's methods is currently limited to about 25 items. Bock developed a factor-analytic approach for dichotomous data, called full-information factor analysis (Bock, Gibbons, and Muraki, 1985) because it uses information contained in the joint frequencies of all orders of the item responses. This method, detailed in section 3.4 below, makes use of the marginal maximum likelihood methods of Bock and Aitkin (1981) for estimating the parameters of the common factor model.

In addition to factor-analytic approaches, a number of other methods of dimensionality assessment have been proposed. For example, Bejar (1980) has recommended comparing the estimated item difficulties (i.e., the estimates of the b_j of equation 1) obtained by calibrating a complete set of test items

to those obtained by performing the calibration separately within content areas. (Bejar [1980] also proposed an additional procedure, which involves computing, for each content area, a scaled score corresponding to each of the two sets of item parameter estimates, and then comparing the results obtained by fitting a one-factor model to each of the two sets of scores.) Although Bejar's (1980) application of the method appeared to yield useful results, Hambleton and Rovinelli (in press) found that the method was unable to discriminate between one-and two-dimensional simulated data sets. Another method that has been proposed is analysis of the residual differences between observed responses and the estimated probabilities of correct responses according to the unidimensional item response model deemed appropriate (e.g., equation 1). Various methods of residual analysis have been proposed; reviews are given by Traub and Wolfe (1981) and Hattie (1985). The rationale is that if the model fits well, the data can be assumed to be consistent with undimensionality. A major drawback is that large residuals may be the result of model violations other than multidimensionality. Hambleton and Rovinelli (in press) concluded that indices based on the size of average residuals obtained after fitting one-, two-, and three-parameter logistic models were not capable of detecting multidimensionality. It should be noted that Hambleton and Rovinelli did not report any investigation of the pattern of residuals.

3. Methods used to assess the dimensionality of NAEP reading data

The proposed methods of dimensionality assessment differ in terms of the assumptions needed, the hypothesis tested, and the statistical artifacts that affect interpretation. Rather than selecting a single method of

dimensionality assessment for the NAEP reading data, we applied four different techniques, described in this section. For descriptive purposes, we included principal components analysis (PCA) of phi and tetrachoric correlations, as described in section 3.2. As an experimental analysis, we also applied PCA to the image correlation matrix, a method based on the work of Guttman (1953) and Kaiser and Cerny (1979), described in section 3.3. Bock's full-information factor analysis, discussed in section 3.4, was applied to a subset of the data. Finally, we used the method of Rosenbaum (1984a, 1984b), described in section 3.5, which involves examination of the partial association for each pair of items, conditional on the total score on the remaining items. Prior to a discussion of these methods, the properties of the NAEP data base are described.

3.1 Properties of NAEP data

3.1.1 Items included in dimensionality analyses

All reading items that were included in the IRT scaling and were also spiralled with other items (see section 3.1.2 and scaling report) were used in the dimensionality analyses. All subjects who responded to one or more of these items were included. The number of subjects and items available for the analyses is shown in Table 1. As indicated, there were about 100 items per grage. Twenty-five of the items included in the analyses were administered to all three grages. The range and mean of the proportions correct for each of the three grages and for the 25 across-grage items are given in Table 1. As shown, the number of students per grage was roughly 26 to 29 thousand, corresponding to weighted frequencies of over 3 million. As a result of the number of items and subjects in the data base, certain analyses were ruled out

Table 1

Number of Items and Students Available for Dimensionality Analyses

Grage	Number of Items	Proportions Correct Minimum	Maximum	Mean	Number of Students Unweighted	Weighted
9/IV	108	.04	.93	.50	26,087	3.5 million
13/VIII	100	.09	.98	.63	28,405	3.3 million
17/IX	95	.21	.96	.70	28,861	3.4 million
Across Grages (Common Items)	25	.13	.90	.53	83,353	10.2 million

because they were too costly or exceeded computing capabilities. In other cases, dimensionality analyses were performed on only a subset of items to minimize the cost and the computational burden.

Ninety-four percent of the NAEP reading items included in the analyses were multiple choice items with three to six response choices. The remainder were essay items in which the respondent was asked to react to a reading passage. Essay items were scored on a scale of 1 to 5, which was later dichotomized. All items were classified by reading experts on the basis of objective (deriving information vs. integrating and applying information), stimulus (short or long reading passage, document, or picture), and content (fictional story, poem, informational passage, social studies, science, arts and humanities, or life skills). These item properties, as well as a further classification of the items based on the work of Mosenthal (1985), were used in attempting to interpret analysis results. (A subset of reading items that were designed to assess study skills were not included in the dimensionality analysis because they were not scaled using IRT. That these items differed from the remaining reading items was suggested by examination of the item content, as well as empirical evidence: For a subset of examinees, number-right scores on blocks of study skills items and on blocks of conventional reading items were obtained. The attenuation-corrected correlations between study skills blocks and conventional reading blocks tended to be lower than intercorrelations between conventional reading blocks. Many of the items which led to departures from unidimensionality in Jungeblut's [1984] analyses of the 1979-1980 NAEP data were study skills items [Jungeblut, personal communication, October, 1985].)

3.1.2 Missing data pattern

A new feature of the year 15 NAEP design was the use of balanced incomplete block (BIB) spiralling to assign test items to booklets (see Messick, Beaton, and Lord, 1983; Beaton, 1984). BIB spiralling combines the features of conventional spiralling and multiple matrix sampling. As in ordinary multiple matrix sampling, each item is administered a prescribed number of times, although examinees receive different subsets of items. BIB spiralling has the additional feature that each pair of items is assessed a prescribed number of times. In NAEP, reading items were first grouped into blocks, consisting in most cases of 8 to 12 items, which were then assigned to test booklets according to a design that conformed to these criteria. This resulted in a set of approximately 60 different test booklets per grage, which were assigned to respondents in a random sequence.

A major advantage of BIB spiralling is that it permits the estimation of inter-item correlations. However, the resulting matrix of correlations, referred to here as the BIB matrix, has an unusual pattern of missing data. In the case of the NAEP reading data, the number of respondents available to estimate correlations between items in the same block is, in most cases, nine times the number of respondents available for the estimation of correlations between items that fall within different blocks. Furthermore, the correlations of items in one block, say, A, with those in another block, B, are not in general based on the same group of respondents as the correlations of Block C items with Block D items. Because of the spiralling procedure used to assign booklets to respondents, the missing data that result from the implementation of a BIB design can be regarded as random. However, in using a BIB correlation matrix rather

than a conventional correlation matrix, we are implicitly making the assumption that the correlations between items are not subject to context effects. If, for example, the population correlation between two items, i and j, varied depending on whether k were administered with i and j, then the sample correlation of i and j in the presence of k would not be an estimate of the same population parameter as the sample correlation of i and j in the absence of k. Computation of a BIB matrix would involve averaging these sample correlations, which would be undesirable.

Even if the assumption of no context effects is justified, there are other ways in which the properties of the BIB matrix differ from those of a conventional correlation matrix. For example, the standard errors of the within-block correlations are smaller than those of the between-block correlations. Also, the BIB matrix may have negative eigenvalues, unlike a conventional correlation matrix. As detailed in section 3.1 and Tables 2 and 3, both phi and tetrachoric matrices of NAEP items had negative roots in most cases. For analyses that required a matrix that was at least positive semi-definite, an adjustment procedure, described in Appendix 1, was applied. Although there is no indication that analysis results were affected in any major way by the use of BIB matrices or their adjusted counterparts, the statistical properties of these matrices are not fully understood at present.

In addition to the BIB missing data, which can be regarded as random, there are two major categories of non-random missing data: omitted items and items that the respondent was administered but did not reach. Unanswered items occurring after the last valid response within a block were considered "not reached." (In administering the items, each block was timed separately.)

Unanswered items that occurred prior to the last valid response (and were not a result of the BIB design) were coded as omits. The category of omitted items was defined to include as well any items marked, "I don't know," which was a response alternative for all multiple choice items. The treatment of not reached and omitted items in each of the dimensionality analyses is discussed in sections 3.2-3.5.

3.2 Principal component analysis of inter-item correlation matrices

Despite the drawbacks described in section 2, principal component analyses (PCA) of the phi and tetrachoric matrices for each grage were conducted for descriptive purposes. In addition, analyses including all respondents were performed, based on the 25 items common to all three grages. It can be argued that the results of these analyses represent a "worst case;" that is, because the analyses tend to produce spurious factors, results that were free of artifacts would be expected to be more consistent with unidimensionality.

Items that were not reached were excluded from the analysis; omitted items were scored as incorrect. For each of the four phi matrices, Table 2 gives the range of inter-item correlations, the median correlation, the first five eigenvalues and the percent of the trace they represent, and, as an index of the degree to which the matrix departed from positive-definiteness, the sum of the negative eigenvalues as a percent of the trace of the matrix. The range of sample sizes (N) on which the correlation coefficients were based (see section 3.1.2) is also given. The corresponding information for the tetrachoric matrices is given in Table 3. The results in Tables 2 and 3 are based on analyses that

incorporated the respondents' sampling weights (see Lago, Burke, Tepping, and Hansen (1985). Unweighted analyses yielded almost identical results.

It is clear that, for each of the eight matrices, there is a large first root, constituting between 17 and 25 percent of the trace for the phi matrices and between 30 and 40 percent for the tetrachoric matrices (but note that the negative roots constitute up to 27 percent of the trace for tetrachoric matrices). The second root is always less than one-fourth of the first. Following the sharp drop-off between the first and the second, the remaining roots trail off gradually. These findings are reassuring in that they are consistent with a large first dimension. (The size of the first component may appear small to those who are unaccustomed to examining the results of item-level factor analyses. In interpreting these findings, however, it is important to consider that the median inter-item correlations are low: between .14 and .19 for the four phi matrices and between .27 and .35 for the tetrachoric matrices. Results of PCA of phi matrices computed from simulated unidimensional data showed that the first root typically constituted 25 to 30 percent of the trace; see section 3.3 and Table 5.) The loadings on the first principal component were not related in any obvious way to the item classifications discussed in section 3.1.1.

3.2.1 Application of guessing corrections to tetrachoric correlations

When it is possible for items to be answered correctly through guessing, the magnitude of observed tetrachoric correlations is related to item difficulty (e.g., see Hulin, Drasgow, and Parsons, 1983,

Table 2

Eigenvalues and Descriptive Statistics for Phi Matrices

Grage 9/IV (108 items)

First 5 Roots	Pct. of trace	Descriptive Statistics	
23.9	22	Range of N	149, 5502
3.3	3		
2.5	2	Range of r	-.18, .53
2.4	2	Median r	.19
2.2	2	Neg. roots as pct. of trace	3

Grage 13/VIII (100 items)

First 5 Roots	Pct. of trace	Descriptive Statistics	
17.0	17	Range of N	160, 4502
2.6	3		
2.5	2	Range of r	-.15, .60
2.2	2	Median r	.14
2.1	2	Neg. roots as pct. of trace	2

Grage 17/IX (95 items)

First 5 Roots	Pct. of trace	Descriptive Statistics	
17.5	18	Range of N	167, 4659
3.1	3		
2.3	2	Range of r	-.16, .68
2.1	2	Median r	.16
2.0	2	Neg. roots as pct. of trace	2

All Grages Combined (25 items)

First 5 Roots	Pct. of trace	Descriptive Statistics	
6.3	25	Range of N	607, 8862
1.5	6		
1.2	5	Range of r	.29, .57
1.1	5	Median r	.18
1.0	4	Neg. roots as pct. of trace	0

Table 3

Eigenvalues and Descriptive Statistics for Tetrachoric Matrices

Grage 9/IV (108 items)

First 5 Roots	Pct. of trace	Descriptive Statistics	
39.5	37	Range of N	149, 5502
6.6	6		
4.7	4	Range of r	-.46, .81
3.7	3	Median r	.35
3.4	3	Neg. roots as pct. of trace	27

Grage 13/VIII (100 items)

First 5 Roots	Pct. of trace	Descriptive Statistics	
30.0	30	Range of N	160, 4502
4.3	4		
3.8	4	Range of r	-.34, .81
3.4	3	Median r	.27
3.3	3	Neg. roots as pct. of trace	21

Grage 17/IX (95 items)

First 5 Roots	Pct. of trace	Descriptive Statistics	
32.0	34	Range of N	167, 4659
3.9	4		
3.3	3	Range of r	-.38, 90
3.0	3	Median r	.31
2.8	3	Neg. roots as pct. of trace	19

All Grages Combined (25 items)

First 5 Roots	Pct. of trace	Descriptive Statistics	
10.0	40	Range of N	607, 8862
1.6	6		
1.2	5	Range of r	.05, .80
1.2	5	Median r	.33
1.0	4	Neg. roots as pct. of trace	0

pp. 249-255). To eliminate this problem, Carroll (1945) suggested that the frequencies in the 2 x 2 tables of responses for each pair of items be adjusted to "remove" the effects of guessing and that tetrachorics be computed on the basis of these adjusted frequencies. In Carroll's model, it is assumed that guessing is random and that the probability of getting an item right by guessing is therefore equal to the reciprocal of the number of response choices. It is also implicitly assumed that, for each pair of items, the probability of getting one item right by guessing is independent of the probability of making a correct guess on the other item. To determine whether it would be a useful strategy for NAEP data, Carroll's correction was applied to the item responses for grage 13/VIII, setting g_j, the hypothetical probability of guessing right on item j, equal to the reciprocal of the number of response choices for item j, excluding the "I don't know" alternative. For essay items, g_j was set to 0. The results were clearly unsatisfactory: It was found that 16 percent of the tetrachoric coefficients were rendered incomputable because of negative adjusted cell frequencies. Several other corrections were investigated, but deemed unsatisfactory, including a modification of Carroll's correction in which the input g_j values were adjusted so as to avoid the occurrence of negative adjusted cell frequencies and a correction in which each g_j was set equal to the estimated lower asymptote, c_j (see equation 1) of the item from the IRT item calibration. Note that Bock, Gibbons, and Muraki [1985] describe a modification of Carroll's correction that apparently produces satisfactory results.)

3.3 Principal components analysis of the image correlation matrix

Guttman (1953) developed a theory for the structure of quantitative variates called image theory. Image theory is based on the partitioning of a variable into two additive segments: the part that can be predicted through least squares linear regression of that variable on all the remaining variables, called the image, and the error of prediction, called the anti-image. Thus, unlike common factor theory, image theory provides an explicit definition for the common part of a variable. Another difference from the traditional factor-analytic approach is that the anti-images may have non-zero covariances. Guttman shows that common factor theory may be viewed as a special case of image theory. The relation between image theory and other factor-analytic approaches is further examined by Harris (1962) and reviewed by Mulaik (1972).

Suppose that n variables are to be observed. The decomposition of the original variates into images and anti-images can be expressed as

$$\underset{\sim}{z} = \underset{\sim}{v} + \underset{\sim}{u} \qquad [2]$$

where $\underset{\sim}{z}$ is the n x 1 vector of observable random variables, standardized to have mean zero and unit variance, $\underset{\sim}{v}$ is the n x 1 vector random variable of images defined in equation 3, below, and $\underset{\sim}{u}$ is the n x 1 vector random variable of anti-images, or errors of prediction. (When referring to a finite sample of variables, Guttman used the terms partial image and partial anti-image. The qualifier, "partial" will not be used here.) The n x 1 vector random variable $\underset{\sim}{v}$ of images can be expressed as

$$\underset{\sim}{v} = \underset{\approx}{W}\underset{\sim}{z} \qquad [3]$$

The weight matrix $\underset{\sim}{W}$ is defined as

$$\underset{\sim}{W} = \underset{\sim}{I} - \underset{\sim}{S}^2 \underset{\sim}{R}^{-1} \qquad [4]$$

where $\underset{\sim}{R}$ is the correlation matrix of the original variates, z, and

$$\underset{\sim}{S}^2 = [\text{diag } (\underset{\sim}{R}^{-1})]^{-1} \qquad [5]$$

The off-diagonals of $\underset{\sim}{W}$ contain the regression weights for predicting each of the variates z from the remaining n - 1 variates. The diagonals of $\underset{\sim}{W}$ are equal to zero because the regression of a variate on itself is not of interest.

The principles of image theory are usually applied in practice by factor-analyzing $\underset{\sim}{G}$, the covariance matrix of the images, given by

$$\begin{aligned} \underset{\sim}{G} &= E(\underset{\sim}{v}\underset{\sim}{v}') = E\,(\underset{\sim}{W}\underset{\sim}{z})\,(\underset{\sim}{W}\underset{\sim}{z})' \qquad [6] \\ &= E(\underset{\sim}{W}\underset{\sim}{z}\underset{\sim}{z}'\underset{\sim}{W}') = \underset{\sim}{W}\,E(\underset{\sim}{z}\underset{\sim}{z}')\,\underset{\sim}{W}' \\ &= \underset{\sim}{W}\underset{\sim}{R}\underset{\sim}{W}' = (\underset{\sim}{I} - \underset{\sim}{S}^2\underset{\sim}{R}^{-1})\,\underset{\sim}{R}\,(\underset{\sim}{I} - \underset{\sim}{S}^2\,\underset{\sim}{R}^{-1})' \\ &= \underset{\sim}{R} + \underset{\sim}{S}^2\,\underset{\sim}{R}^{-1}\,\underset{\sim}{S}^2 - 2\underset{\sim}{S}^2 \end{aligned}$$

The j^{th} diagonal element of this matrix is the variance of the j^{th} image, which is equal to the squared multiple correlation coefficient (SMC) obtained by regressing the j^{th} variate on the remaining n - 1 variates. In this sense, $\underset{\sim}{G}$ resembles the "reduced correlation matrix" of common factor analysis with SMCs used as communality estimates. The off-diagonals of $\underset{\sim}{G}$, however, tend to be slightly smaller than those of the reduced correlation matrix (Kaiser,

1963); furthermore, $\underset{\sim}{G}$ is always Gramian (assuming data are complete), unlike a correlation matrix with SMCs inserted in the diagonal.

As an alternative to the analysis of the $\underset{\sim}{G}$ matrix, Kaiser and Cerny (1979) recommended principal component analysis of the image correlation matrix, $\underset{\sim}{G}^*$, given by

$$\underset{\sim}{G}^* = \underset{\sim}{D}^{-1/2} \underset{\sim}{G} \underset{\sim}{D}^{-1/2} \qquad [7]$$

where

$$\underset{\sim}{D} = \text{diag}\ (\underset{\sim}{G}) = \underset{\sim}{I} - \underset{\sim}{S}^2 \qquad [8]$$

Kaiser (1970; see also Kaiser and Cerny, 1979) conjectured that image analysis would be well-suited to the factor analysis of dichotomous data. He noted that because the images are least squares predicted values of one variate based on the remaining n - 1 variates, "a crude appeal to the Central Limit Theorem suggests that the images will be sensibly multivariate normal, a set-up which is well known not to produce difficulty factors" (Kaiser, 1970, p. 407. Although McDonald and Ahlawat (1974) expressed doubt about the utility of this approach, some unpublished work by Meredith (personal communication, September, 1985) provided partial confirmation of Kaiser's conjecture.

As an experimental approach to dimensionality assessment, principal component analysis of the image correlation matrix was applied to the NAEP data for grages 9/IV, 13/VIII, and 17/IX and to the 25 across-grage items. Modification of the standard equations of image analysis was required because, in the case of NAEP data, the matrix $\underset{\sim}{R}$ of weighted phi correlations is not

positive definite (see section 3.2 and Table 2) and therefore can not be inverted. An adjustment procedure, detailed in Appendix 1, was used to obtain a singular approximation to the matrix of inter-item correlations and a pseudo-inverse of this adjusted matrix. Following this, the pseudo-inverse matrix $\underset{\sim}{R}^{-}$ was then substituted for $\underset{\sim}{R}^{-1}$ in the formulas for $\underset{\sim}{W}$ and $\underset{\sim}{S}^2$ (equations 3 and 4), as recommended by Kaiser and Cerny (1978). Analogues of the matrices $\underset{\sim}{G}$, $\underset{\sim}{G}^*$, and $\underset{\sim}{D}$ (equations 6, 7, and 8) were computed using these modified forms of $\underset{\sim}{W}$ and $\underset{\sim}{S}^2$.

The first five roots of the image correlation matrix are given in Table 4 for the three grages and for the across-grage analysis. For the three within-grage analyses, the first roots are between 14 and 47 percent larger than those for the Pearson matrix. There are at least two possible reasons for this. One distinction between the two PCA methods, which applies regardless of whether the data are dichotomous, is that the PCA of the Pearson matrix involves the correlations of observed values on the original variates z (equation 2), whereas PCA of the $\underset{\sim}{G}^*$ matrix involves the correlations of the common parts, v, of the items as defined in equations 2-5. This difference would be expected to result in larger first roots for the image approach. Furthermore, in the present application of image analysis, the problems associated with linear factor analysis of dichotomous data are to some degree ameliorated by using a matrix of correlations between weighted sums of dichotomous item scores. This, of course, was the basis for Kaiser's conjecture that the image approach would work well in the dichotomous case. It is somewhat surprising that the second roots are also substantially larger for the image matrix than for the Pearson matrix. This is most obvious in

Table 4

Eigenvalues of the Image Correlation Matrix

Grage 9/IV (108 items)

First 5 Roots	Pct. of trace
27.3	25
9.5	9
3.7	3
3.2	3
2.7	3

Grage 13/VIII (100 items)

First 5 Roots	Pct. of trace
23.2	23
9.5	9
3.9	4
2.8	3
2.6	3

Grage 17/IX (95 items)

First 5 Roots	Pct. of trace
25.8	27
5.7	6
4.3	4
3.4	4
3.3	3

All Grages Combined (25 items)

First 5 Roots	Pct. of trace
18.0	72
2.0	8
1.1	5
0.7	3
0.6	2

grage 13/VIII, where the second root of the image correlation matrix is more than three times as large as the second root of the Pearson matrix.

For each of the three within-grage analyses, a solution conforming to the principles of simple structure could be obtained using Promax rotation (Hendrickson and White, 1964). In order to interpret the factors, the relation between the loadings for the rotated solutions and the classifications of reading items described in section 3.1.1 was examined. No clear pattern emerged, however. Furthermore, for items that were administered to more than one grage, there was no consistency across grages in the configurations of loadings.

Results for the 25 items that were administered to all three grages were substantially different from the within-grage analyses. The first root of the image correlation matrix constituted more than seventy percent of the trace, a finding that appears consistent with unidimensionality. The first root was nearly three times the size of the first root of the Pearson matrix; the second root grew only slightly in this case. It is likely that results of this analysis differed from those of the within-grage analysis because the across-grage correlation matrix was better-behaved. The sample sizes were larger and there were no negative correlations or negative roots.

To aid in interpreting the results of the four image analyses, PCA of the image correlation matrix was applied to several simulated data sets generated from a unidimensional model. The simulation studies were conducted as follows: (1) Assuming a three-parameter logistic model, NAEP reading items were calibrated with the LOGIST program (M. S. Wingersky, 1983) using actual NAEP data. Thirty of these items were randomly selected for this simulation run. (2) One thousand pseudo-random values from a normal distribution with mean zero and unit variance were then generated. These represent theta or proficiency values for N = 1000 examinees. (3) For each examinee, the three-parameter logistic function (equation 1) was used to obtain the n x N = 30 x 1000 values of P_{ij}, the probability that person i gets item j correct. The item parameters a_j, b_j, and c_j, were obtained from step 1 and the θ_i values from step 2. (4) Corresponding to each value of P_{ij}, a pseudo-random value U_{ij} was generated from a uniform distribution on the interval [0,1]. If U_{ij} was less than P_{ij}, item j was scored as correct for person i; otherwise it was scored as incorrect. The correlation matrix of these simulated data was then obtained and the image procedure applied.

Table 5 shows the first five roots of the phi and image correlation matrices for one of the simulated data sets. Results were much more dramatic than for the within-grage analyses of the actual NAEP data; the findings bore a closer resemblance to the across-grage analysis of 25 items. Whereas the first root of the phi matrix was only about one quarter of the trace in the simulation, the first root of the image correlation matrix was about 80 percent of the trace. Other simulated unidimensional data sets produced similar values. If the size of the first root is used as a criterion, the image analysis technique is superior

Table 5

First Five Eigenvalues of Correlation and Image Correlation Matrices for Simulation Data (30 items with NAEP item parameters)

Phi Matrix		Image Correlation Matrix	
First 5 Roots	Pct. of Trace	First 5 Roots	Pct. of Trace
7.7	26	23.8	79
1.7	6	2.6	9
1.1	4	0.5	2
1.0	3	0.5	2
1.0	3	0.4	1

Correlation of Loadings on Second Principal Component with Proportions Correct

.85 .65

to PCA of the phi matrix in revealing the true unidimensional structure underlying the data. However, as in the case of the phi matrix, the loadings of items on the second principal component of the image correlation matrix have substantial correlations with the proportions correct for the items: the correlations were .85 for the phi matrix and .65 for the image correlation matrix. Because it is evident that the results of the PCA of the image correlation matrix are not free of statistical artifacts, no further attempt was made to interpret the Promax solutions. (It should also be noted that no simulation studies of the performance of image analysis under multidimensionality were conducted.)

3.4 Bock's full-information factor analysis

Another factor-analytic method that was applied to the NAEP data is Bock's full-information factor analysis (Bock, Gibbons, and Muraki, 1985; see also Mislevy, in press), which is implemented in the TESTFACT program (Wilson, Wood, and Gibbons, 1983). Unlike the methods described in sections 3.2 and 3.3, this method does not require the computation of correlation coefficents, but operates instead on the n-way contingency table of item responses. In contrast to factor analysis of correlation coefficients, which makes use of only the pairwise joint frequencies of item responses, Bock's full-information solution uses information contained in the joint frequencies of all orders. In applying this method, a particular model for the item responses must be assumed. In the case of the NAEP data, the selected model was a multivariate generalization of the three-parameter normal ogive in which each item is allowed to load on multiple factors. The model can be developed by first

assuming that underlying the response of person i to item j is a response process variable defined as

$$y_{ij} = \sum_{k=1}^{K} \lambda_{jk}\, \theta_{ki} + \nu_j \qquad [9]$$

where θ_{ki} represents the value of the k^{th} latent variable (factor), k = 1, 2, ... K, for the i^{th} individual, i = 1, 2, ... N, λ_{jk} is the loading of the j^{th} item, j = 1, 2, ... n, on the k^{th} latent variable, and ν_j is a residual term associated with item j. The observed score of the i^{th} examinee on the j^{th} item, x_{ij}, takes on a value of 1, indicating a correct score, if y_{ij} exceeds γ_j, the threshold for the j^{th} item. If it is assumed that the residuals ν_j are independently distributed as N(0, σ_j), the conditional probability that the i^{th} examinee gets the j^{th} item correct, given that his values on the latent variable are equal to $\underline{\theta}_i$ can be expressed as

$$P(x_{ij} = 1 \mid \underline{\theta}_i) = \frac{1}{\sqrt{2\pi}\sigma_j} \int_{\gamma_j}^{\infty} \exp\left[-1/2 \left(\frac{y - \sum_{k=1}^{K} \lambda_{jk}\theta_{ki}}{\sigma_j}\right)^2\right] dy \qquad [10]$$

$$\equiv F_j(\underline{\theta}_i)$$

This is a multivariate generalization of the two-parameter normal ogive model (see Lord and Novick, 1968).

This model can be modified to allow for the possibility of guessing by substituting

$$F_j^*(\underline{\theta}_i) = c_j + (1 - c_j) F_j(\underline{\theta}_i) \qquad [11]$$

for $F_j(\underline{\theta}_i)$, where c_j represents the probability that an individual with very low ability gets the item correct. This multivariate generalization of the three-parameter normal ogive model was applied in the NAEP analyses. The c_j parameters were estimated using BILOG (Mislevy and Bock, 1982) and then input to the TESTFACT program. NAEP items that were coded as "not reached" (see section 3.1.2) were not included in the analysis. Omitted items, on the other hand, were scored correct with probability c_j. Under this strategy, examinees who omit an item have the same theoretical probability of getting the item correct as examinees who guess in the absence of any information.

Incorporating the item response function, $F_j^*(\underline{\theta}_i)$, defined in Equation 11, the marginal probability of the s^{th} response pattern can be expressed as:

$$P_s = P(\underline{x} = \underline{x}_s) = \int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} \prod_{j=1}^{n} F_j^*(\underline{\theta})^{x_{sj}} [1 - F_j^*(\underline{\theta})]^{1-x_{sj}} f(\underline{\theta}) d\underline{\theta} \qquad [12]$$

where x_{sj} is the response to the j^{th} item in the s^{th} response pattern, s = 1, 2, ... S, and $S \leq \min(2^n, N)$ is the number of response patterns. It is further assumed in this application that $f(\underline{\theta})$ is the multivariate normal distribution with mean $\underline{0}$ and covariance matrix $\underline{I}$. Now, if it is assumed that the counts of the distinct response patterns follow a multinomial distribution, the likelihood of the matrix $\underline{X}$ of observed counts r_s of distinct response patterns can be expressed as:

$$P(\underline{X}) = \frac{N!}{r_1! r_2! \ldots r_s!} P_1^{r_1} P_2^{r_2} \ldots P_s^{r_s} \qquad [13]$$

where P_s is given by Equation 12.

The quantities P_s are estimated using numerical integration techniques. The marginal maximum likelihood method of Bock and Aitkin (1981), which is based on earlier work by Bock and Lieberman (1970), is then applied to Equation 13 to obtain estimates of the factor loadings and thresholds for each item (see Bock, Gibbons, and Muraki, 1985; Mislevy, in press).

If sample size is sufficiently large, a test of the fit of the K-factor model relative to a general multinomial alternative can be obtained using a chi-square approximation to the likelihood ratio test. The model can be re-estimated and the test repeated for successive values of K. The difference between these chi-square statistics is also distributed as chi-square (under the hypothesis that the more restrictive model is correct) and can be used to test the improvement in model fit that is achieved by allowing the number of latent variables to increase. The test of change in model fit has been shown to perform well even when the frequency table is sparse (Haberman, 1977).

Because the TESTFACT program is very expensive to run, full-information factor analysis was applied only to 42 items for grage 13/VIII. These items, which were chosen to maximize the chances of detecting multidimensionality, were intended to represent four distinct item types: reading comprehension, vocabulary, life skills and essay. The comprehension, vocabulary, and essay items all referred to passages the examinee was asked to read. Some passages were fictional stories; others pertained to an academic content area, such as science or social studies. The life skills items were based on documents that might be encountered in everyday life, such as a portion of a telephone directory, a grocery store coupon, or an advertisement.

The analysis was based on the raw rather than the weighted frequency table of item responses. Because sampling weights have little effect on variances and covariances, they are unlikely to have much effect on factor analysis results (Bock, personal communication, November, 1985).

In applying the chi-square test for the number of latent variables or factors, it was necessary to take into account the effects of multistage cluster sampling (see Lago et al., 1985) on the variability of the test statistic. In adjusting the significance tests, it was assumed that the design effect was equal to two. Research conducted with previous NAEP surveys led to the conclusion that this was a reasonable estimate of the design effect for this type of test statistic (Johnson, 1980). This means that an estimate of the variability of the test statistic under the NAEP sample design can be obtained by computing the variance of the statistic under simple random sampling assumptions and then multiplying the obtained value by two. Because the log likelihood chi-square statistics are proportional to sample size, a design effect can be incorporated

simply by dividing the chi-square values by the design effect. Incorporating this adjustment, the chi-square test corresponding to the change from the one- to the two-factor solution was not significant, indicating that the one-factor solution could be retained. The single factor accounted for about 39 percent of the total variance. Reading comprehension items, particularly those that involved fictional stories, tended to have the highest factor loadings. Life skills items had the lowest loadings.

3.5 Rosenbaum's test of unidimensionality, monotonicity, and conditional independence

Rosenbaum (1984a) proves a theorem that states that if item characteristic curves are nondecreasing functions of a single latent variable, then conditional (local) independence of item responses, given the latent variable, implies certain relations among the item responses. Specifically, the conditional covariances between all monotone increasing functions of a set of item responses, given any function of the remaining item responses, will be non-negative. This theorem can be used to develop statistical tests of whether an observed data set is consistent with the assumptions of monotonicity, unidimensionality, and conditional independence. (See Holland, 1981, Holland and Rosenbaum, in press, and Stout, 1984, for further discussion of tests of this kind.)

As a special case of Rosenbaum's theorem, we can test the partial association for each pair of items, given number-right score on the remaining

items, using the Mantel-Haenszel (1959) test, a conventional procedure for analysis of discrete data. In this case, we are examining the conditional covariance between monotone item summaries which are simply responses to a single item. The function on which we are conditioning is the number-right score on the remaining n - 2 items. To perform the Mantel-Haenszel test for a particular item pair, a 2 x 2 table of item responses is constructed for each of the K possible values of number-right score on the remaining items. Let n_{ijk} be the observed count in the i^{th} row, jth column, and k^{th} table, where $i = 1, 0$; $j = 1, 0$; and $k = 1, 2, \ldots K$. The Mantel-Haenszel test statistic is given by

$$Z = \frac{n_{11+} - E(n_{11+}) + 1/2}{\sqrt{V\ (n_{11+})}} \qquad [14]$$

where $E(n_{11+})$ and $V(n_{11+})$ denote the hypergeometric expectation and variance of n(11+), given by

$$E(n_{11+}) = \sum_{k=1}^{K} \frac{n_{1+k}\ n_{+1k}}{n_{++k}} \qquad [15]$$

$$V(n_{11+}) = \sum_{k=1}^{K} \frac{n_{1+k}\ n_{0+k}\ n_{+1k} n_{+0k}}{n^2_{++k}\ (n_{++k} - 1)} \qquad [16]$$

and the plus subscript indicates summation over that subscript. The approximate significance level is obtained by referring Z to the lower tail of the standard normal distribution. A statistically significant result indicates that the pair of items has a negative partial association and is thus inconsistent with the hypothesized model.

The Mantel-Haenszel approach was programmed to accommodate the complexities of BIB spiralling in the following way: Suppose that we are interested in assessing the conditional covariance between items X_1 and X_2 and that, because of BIB spiralling, certain students who received items X_1 and X_2 also received X_3, X_4, and X_5, whereas others received X_5 and X_6. The test of association between X_1 and X_2 is then based on seven 2 x 2 tables: four corresponding to the possible score values for $X_3 + X_4 + X_5$ and three for the possible scores for $X_5 + X_6$. Because of the spiralling method used to assign booklets to respondents (see section 3.1.2), the fact that respondents did not all receive the same items or even the same number of items does not impair the validity of the method. Items that were omitted or were administered but not reached (see section 3.1.2) were scored as incorrect.

Because of the cost of computations, the Rosenbaum method was applied to only a subset of the NAEP items: those in blocks H, K, M, N, and O. The number of items per grage was 56 for grage 9/IV, 53 for grage 13/VIII, and 55 for grage 17/IX. The number of hypothesis tests, which is equal to the number of item pairs, was 1540, 1378, and 1485 for grages 9/IV, 13/VIII, and 17/IX respectively. In order to evaluate the findings of this method, a decision must be made about the appropriate alpha level at which to test these multiple hypotheses. Whereas on one hand, we would like to control the overall Type I error rate at an acceptable level, we do not want to maintain such rigorous Type I error control that a rejection of the hypothesis of unidimensionality would be impossible. As it turns out, even if the alpha for each hypothesis test is set at .01, a liberal alpha level for so large a number of tests, the number of statistically significant negative partial associations is only 4

Table 6

Results of Rosenbaum Analyses

Within-Grage Anayses

	Grage		
	9/IV	13/VIII	17/IX
Number of items	56	53	55
Number of item pairs	1540	1378	1485
Number of significant negative partial associations:			
α = .01 per comparison	4	4	6
α = .05 per comparison	31	29	26

Across-Grage Analyses

	Grage pair		
	9 & 13	9 & 17	13 & 17
Number of comparisons	24	24	24
Number of significant negative partial associations:			
α = .05 per comparison	0	0	0

for grage 9/IV, 4 for grage 13/VIII, and 6 for grage 17/IX. If alpha is set at .05 for each test, the number of statistically significant results is 31, 29, and 26 for the three grages, respectively (see Table 6). Therefore, it is reasonable to retain the hypothesis that the item responses can be represented by a monotonic unidimensional latent variable model with conditional independence. It should be noted that application of the Rosenbaum method does not provide a test of the fit of the three-parameter logistic model or of any other specific model.

In applying the Rosenbaum method, no modifications were incorporated to reflect NAEP's complex multistage cluster sampling scheme (Lago et al., 1985). That is, raw rather than weighted frequencies were used in the analysis and no jackknifing or design effect adjustment was used in computing the significance probabilities of the Mantel-Haenszel statistics. As noted in section 3.2, weighted and unweighted correlation matrices for the NAEP data are virtually identical, suggesting that the weights would make little difference in the Rosenbaum analyses. Furthermore, the design effect for these tests is likely to be greater than one, as in 3.4. Adjustment of the significance tests would then lead to a reduction in the number of item pairs found to have negative partial associations, thus reinforcing the original conclusion about dimensionality.

3.5.1 Across-grage analyses

In addition to determining whether it was reasonable to regard the reading items as unidimensional within each grage, it was of interest to investigate whether unidimensionality would hold if respondents

from all three grages were included. Of the entire set of items available for dimensionality analyses (Table 1), 25 were administered to all three grages. Twenty-four of these 25 were in the item blocks (H, K, M, N, O) used for the Rosenbaum analyses. A method developed by Rosenbaum (1984b), which is a variant of the approach described above, was applied to these 24 items. The procedure provides a test of whether the item responses of two groups of examinees is consistent with a difference in the distribution of a unidimensional latent variable. A rejection of this hypothesis would mean that it was necessary to postulate the existence of additional dimensions. As a first step in the analysis, an indicator variable is created to represent group membership, with the higher value associated with the group hypothesized to have generally higher values on the latent variable. If the pattern of item responses is consistent with the hypothesized model, the conditional covariances of each item with the indicator variable will be non-negative, as described in 3.5.

For the NAEP data, a separate analysis was conducted for each pair of grages, as follows: An indicator variable representing grage was created, with a value of 1 indicating the higher grage and the value of 0 corresponding to the lower grage. The partial association of each of the 24 items with grage was then assessed, using the Mantel-Haenszel (1959) test, as described in 3.5. With an alpha of .05 for each of the 24 hypothesis tests per grage (see Table 6), no significant negative partial associations of items with the dummy-coded grage variable were found. This means that, as we would expect intuitively, students in higher grages were more likely than students in lower grages to answer items correctly, conditional on number-right score on the

remaining items. These results are consistent with unidimensionality of the item pool.

4. Conclusions

Overall, the four dimensionality analyses of the NAEP reading items indicate that it is not unreasonable to treat the data as undimensional. As a preliminary approach, principal component analyses of phi and tetrachoric correlation matrices were computed for each of the three grages and for the 25 across-grage items. The first roots obtained from these analyses were sizeable, ranging from 17 to 25 percent of the trace for the phi matrices and 30 to 40 percent for the tetrachoric matrices. (For simulated unidimensional data, the first root of the phi matrix typically constituted 25 to 30 percent of the trace.)

As an experimental method, a factor-analytic approach based on Guttman's image theory was also applied. Principal component analysis of the image correlation matrices yielded larger first roots than PCA of the corresponding phi matrices, but larger second roots as well. Application of image analysis to simulated unidimensional data showed that principal component loadings had a substantial correlation with the proportions correct for the items. Thus, the image approach does not avoid the artifacts associated with the application of linear factor-analytic methods to dichotomous data.

Application of Bock's full-information factor analysis to a subset of the grage 13/VIII data led to a satisfactory fit with a one-factor model. The first factor accounted for 29 percent of the total variance. Reading comprehension items involving fictional stories had the highest loadings on this factor; life skills items had the lowest.

Finally, the Mantel-Haenszel approach developed by Rosenbaum led to a retention of the hypothesis that the data can be represented by a unidimensional latent variable model with conditional independence. In addition to analyses within each grage, tests were conducted to determine whether data for each pair of grages were consistent with a difference in distribution of a unidimensional latent variable. Again, the hypothesis of unidimensionality was retained.

Although categorization of the NAEP reading items is useful for test development and reading research, the dimensionality analyses reported here do not provide strong empirical evidence for the existence of multiple dimensions. Especially when considered in light of the robustness research discussed in section 1.1, the results do not contraindicate the application of unidimensional item response theory models to the reading data.

References

Beaton, A. E. (1984). Statistical issues in data analysis for the National Assessment of Educational Progress. Paper presented at the annual meeting of the American Statistical Association, Philadelphia, August 1984.

Bejar, I. I. (1980). A procedure for investigating the unidimensionality of achievement tests based on item parameter estimates. Journal of Educational Measurement, 17, 283-296.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Bock, R. D., & Aitkin, M. (1981). Marginal maximum likelihood estimation of item parameters: Application of an EM algorithm. Psychometrika, 46, 443-459.

Bock, R. D., Gibbons, R. D., and Muraki, E. (1985). Full-information item factor analysis (MRC Report No. 85-1). Chicago: National Opinion Research Center.

Bock, R. D., & Lieberman, M. (1970). Fitting a response model for n dichotomously scored items. Psychometrika, 35, 179-197.

Carroll, J. B. (1945). The effect of difficulty and chance success on correlations between items and between tests. Psychometrika, 26, 347-372.

Carroll, J. B. (1983). The difficulty of a test and its factor composition revisited. In H. Wainer & S. Messick (Eds.), Principals of modern psychological measurement. Hillsdale, NJ: Erlbaum.

Christoffersson, A. (1975). Factor analysis of dichotomized variables. Psychometrika, 40, 5-32.

Cook, L. L., Dorans, N. J., Eignor, D. R., & Petersen, N. S. (1985). An assessment of the relationship between the assumption of unidimensionality and the quality of IRT true-score equating. (ETS Research Report 85-30.) Princeton, NJ: Educational Testing Service.

Cook, L. L., & Eignor, D. R. (1984). Assessing the dimensionality of NAEP reading test items: Confirmatory factor analysis of item parcel data. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, April, 1984.

Drasgow, F., & Parsons, C. K. (1983). Application of unidimensional item response theory models to multidimensional data. Applied Psychological Measurement, 7, 189-199.

Guttman, L. (1953) Image theory for the structure of quantitative variates. Psychometrika, 18, 277-296.

Haberman, J. S. (1977). Log-linear models and frequency tables with small expected cells counts. Annals of Statistics, 5, 1148-1169.

Hambleton, R. K., & Rovinelli, R. J. (in press). Assessing the dimensionality of a set of test items. Applied Psychological Measurement.

Harris, C. W. (1962). Some Rao-Guttman relationships. Psychometrika, 27, 247-263.

Hattie, J. (1984). An empirical study of various indices for determining unidimensionality. Multivariate Behavioral Research, 19, 49-78.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. Applied Psychological Measurement, 9, 139-164.

Hendrickson, A. E., & White, P. O. (1964). PROMAX: A quick method for rotation to oblique simple structure. British Journal of Statistical Psychology, 17, 65-70.

Holland, P. W. (1981). When are item response models consistent with observed data? Psychometrika, 46, 79-92.

Holland, P. W., & Rosenbaum, P. R. (In press). Conditional association and unidimensionality in monotone latent variable models. Annals of Statistics.

Hulin, C. L., Drasgow, F., & Parsons, C. K. (1983). Item response theory: Application to psychological measurement. Homewood, IL.: Dow Jones-Irwin.

Johnson, E. G. (1980). Analysis of NAEP data. Technical report, Education Commission of the States, Denver, Colorado.

Jungeblut, A. (1984). Assessing the dimensionality of NAEP reading test items: Linear factor analysis models. Paper presented at the annual meeting of the American Educational Research Association, New Orleans, April, 1984.

Kaiser, H. F. (1963). Image analysis. In C. W. Harris (ed.), Problems in measuring change, pp. 156-166. Madison, WI: University of Wisconsin Press.

Kaiser, H. F. (1970). A second generation little jiffy. Psychometrika, 35, 401-415.

Kaiser, H. F. & Cerny, B. A. (1978). Pseudo-images and pseudo-anti-images from the pseudo-inverse of a singular correlation matrix. British Journal of Statistical Psychology, 31, 99-101.

Kaiser, H. F., & Cerny, B. A. (1979). Factor analysis of the image correlation matrix. Educational and Psychological Measurement, 39, 711-714.

Kingston, N. M., & Dorans, N. J. (1981). The feasibility of using item response theory as a psychometric model for the GRE Aptitude Test. (GRE Board Professional Report 79-12.) Princeton, NJ: Educational Testing Service.

Lago, J. A., Burke, J. S., Tepping, B. J., & Hansen, M. H. (1985). Report on sample selection, weighting, and variance estimation: NAEP-Year 15. Rockville, MD: Westat.

Lord, F. M. (1980). Applications of item response theory to practical testing problems. Hillsdale, NJ: Erlbaum.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the retrospective study of disease. Journal of the National Cancer Institute, 22, 719-748.

McDonald, R. P. (1967). Nonlinear factor analysis. Psychometric Monographs (No. 15.)

McDonald, R. P. (1983). Exploratory and confirmatory nonlinear common factor analysis. In H. Wainer & S. Messick (Eds.). Principals of modern psychological measurement. Hillsdale, NJ: Erlbaum.

McDonald, R. P., & Ahlawat, K. S. (1974). Difficulty factors in binary data. British Journal of Mathematical and Statistical Psychology, 27, 82-99.

Messick, S., Beaton, A. & Lord, F. (1983). NAEP reconsidered: A new design for a new era. (NAEP Report 83-1.) Princeton, NJ: Educational Testing Service.

Mislevy, R. J. (in press). Recent developments in the factor analysis of categorical variables. Journal of Educational Statistics.

Mislevy, R. J., & Bock, R. D. (1982). BILOG: Item analysis and test scoring with binary logistic models [Computer program]. Mooresville, IN: Scientific Software.

Mosenthal, P. B. (1985). An analysis of NAEP reading assessment items. Unpublished manuscript, Syracuse University.

Mulaik, S. A. (1972). The foundations of factor analysis. New York: McGraw-Hill.

Muthén, B. (1978). Contributions to factor analysis of dichotomous variables. Psychometrika, 43, 551-560.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. Journal of Educational Statistics, 4, 207-230.

Rosenbaum, P. R. (1984). Testing the conditional independence and monotonicity assumptions of item response theory. Psychometrika, 49, 425-435. (a)

Rosenbaum, P. R. (1984). Are the item responses of two groups of examinees consistent with a difference in the distribution of a unidimensional latent variable? (Program Statstics Research Technical Report No. 84-51). Princeton, NJ: Educational Testing Service. (b)

Stout, W. F. (1984). The statistical assessment of latent trait dimensionality in psychological testing. (ONR Report). Urbana-Champaign, IL: Department of Mathematics, University of Illinois.

Traub, R E, & Wolfe, R. G. (1981). Latent trait theories and the assessment of educational achievement. Review of Research in Education, 9, 377-435.

Wingersky, B. (1984). Gramianizing matrices. Unpublished memorandum.

Wingersky, M. S. (1983). LOGIST: A program for computing maximum likelihood procedures for logistic test models. In R. Hambleton (ed.), Applications of item response theory. Vancouver, BC: Educational Research Institute of British Columbia.

Wilson, D., Wood, R. L., & Gibbons, R. (1983). TESTFACT: Test scoring and item factor analysis [Computer program.] Chicago: Scientific Software.

Appendix 1

A Procedure for Obtaining a Gramian Matrix that Approximates a BIB Correlation Matrix for NAEP Items

1. Start with the weighted (i.e., incorporating sampling weights) BIB covariance matrix.

2. Substitute zeroes for the negative eigenvalues. (The negative eigenvalues constituted 4, 2, and 2 percent of the trace of the missing data covariance matrix for grages 9/IV, 13/VIII, and 17/XI, respectively. There were no negative eigenvalues for the across-grage matrix.)

3. Now obtain the "reconstructed" covariance matrix, $\underset{\sim}{C}^*$, using the following equation:

$$\underset{\sim}{C}^* = \underset{\sim}{Q}\ \underset{\sim}{D}^*\ \underset{\sim}{Q}',$$

where $\underset{\sim}{Q}$ is the matrix of normalized eigenvectors of the original covariance matrix and $\underset{\sim}{D}^*$ is a diagonal matrix of eigenvalues, with zeroes substituted for the negative eigenvalues. $\underset{\sim}{C}^{*-} = \underset{\sim}{Q}\ \underset{\sim}{D}^{*-1}\underset{\sim}{Q}'$ is the pseudo-inverse of $\underset{\sim}{C}^*$, where the elements of $\underset{\sim}{D}^{*\ -1}$ are the reciprocals of the corresponding elements of $\underset{\sim}{D}^*$ for positive elements of $\underset{\sim}{D}^*$ and zeroes for zero elements of $\underset{\sim}{D}^*$.

4. It is now possible to obtain a reconstructed correlation matrix, $\underset{\sim}{R}^*$, corresponding to $\underset{\sim}{C}^*$, using ordinary methods. The pseudo-inverse of $\underset{\sim}{R}^*$ can be obtained as follows:

$$\underset{\sim}{R}^{*-} = \underset{\sim}{S}\ \underset{\sim}{C}^{*-}\underset{\sim}{S},$$

where $\underset{\sim}{S}$ is a diagonal matrix of the square roots of the diagonal elements of $\underset{\sim}{C}^*$.

It is desirable to begin with the covariance matrix in Step 1 because operating on the correlation matrix, $\underset{\sim}{R}$, directly will produce a reconstructed $\underset{\sim}{R}$ that does not have ones on the diagonal.

The medians of the residuals obtained by subtracting elements of $\underset{\sim}{R}^*$ from elements of the original $\underset{\sim}{R}$ were .007, .002, and .003 for grages 9/IV, 13/VIII, and 17/IX, respectively. In addition, the eigenstructures for the $\underset{\sim}{R}^*$ matrices were very similar to those for the original $\underset{\sim}{R}$'s. The method is inexpensive and is not difficult to program. An alternative method of B. Wingersky (1984) produced smaller residuals, but was prohibitively expensive to execute.